DENOISING OF HUMAN SPEECH USING COMBINED ACOUSTIC AND EM SENSOR SIGNAL PROCESSING

Ng, L. C.; Burnett, G. C.; Holzrichter, J. F.; and Gable, T. J.
Lawrence Livermore National Laboratory and University of California, Davis
P.O. Box 808, L-3
Livermore, California 94550 USA

ABSTRACT

Low Power EM radar-like sensors have made it possible to measure properties of the human speech production system in real-time, without acoustic interference. This greatly enhances the quality and quantify of information for many speech related applications. See Holzrichter, Burnett, Ng, and Lea, J. Acoustic. Soc. Am. 103 (1) 622 (1998). By using combined Glottal-EM-Sensor- and Acoustic-signals, segments of voiced, unvoiced, and no-speech can be reliably defined. Real-time Denoising filters can be constructed to remove noise from the user's corresponding speech signal.

1. INTRODUCTION

Acoustic speech signals carry a great deal of information that can be automatically converted to text, coded for transmission, and many other applications. However, under conditions with a great deal of background noise, with speakers who do not speak clearly (e.g., who co-articulate, or incompletely articulate, etc.) or who speak with strong accents, such systems often do not work adequately. Many mechanisms, by which additional information, describing conditions of the vocal articulators as the speech signal is generated, have been examined to increase the accuracy of automated systems. Examples are TV images of the lip opening, jaw open-close sensors, electro-glottalgraph signals of the vocal fold conditions, etc.

Recently, it has been shown that very low power Electro Magnetic (EM) radar-like sensors can measure conditions of many of the internal (and external) vocal articulators and vocal tract parameters, in real-time, as speech is generated, Holzrichter (1). In particular, a voiced excitation function of speech has been obtained by associating EM sensor signals from the glottal region (i.e., Glottal Electro Magnetic Sensors, or GEMS) with sub- or supra-glottal air pressure pulsations, Burnett (2). These data, combined with corresponding acoustic data, enable robust methods for sampling background noise data, and vastly increase the quality and quantity of information for almost all applications involving speech processing and use.

In addition, these techniques enable accurate definitions of time periods of phonation, and using the statistics of the user's language (3) enable the definition of periods preceding and following phonation when unvoiced speech is likely to occur. In addition, they enable the determination of periods of no speech, when sampling of background noise signals can reliably take place. Along with robust speech presence determination, the timing and spectral content of the determined excitation function enable real-time filters to be constructed for purposes of denoising corresponding acoustic signal segments.

2. HOMODYNE SENSORS

EM radar-like sensors have been designed to transmit EM waves at 2.3 GHz with 0.2 mW of total power. This level is well below continuous international exposure standards for human use. The sensors use a homodyne field disturbance mode of operation that resembles an interferometer measuring the reflection of a transmitted wave against a local (phase reference) wave. As a reflecting interface moves, the phase of the reflected wave varies with respect to the stationary local wave, and a signal associated with this change is detected by a mixer and filter combination. The EM sensor positioned near the glottis in Fig. 1 measures the positional changes of the rear tracheal wall surface, as the air/tissue interface moves versus time, driven by air pressure waves from the glottis opening and closing.

By estimating the EM wave path from the antenna through a high dielectric medium, such as human neck tissue where $\varepsilon = 50$, and across the human trachea, the expected signal from a moving, reflecting air/tracheal-wall interface was obtained [2]. Using the well-known homodyne-radar sensitivity function, Burnett noted that the rear trachea wall behaved as if it were at a 75-mm distance in air. This equivalent distance in tissue plus air tube is consistant with dimensions obtained from CTR images of his neck. This tracheal wall motion signal (i.e., "ballooning") is deconvolved from the wall-tissue response function to obtain a pressure versus time excitation signal.

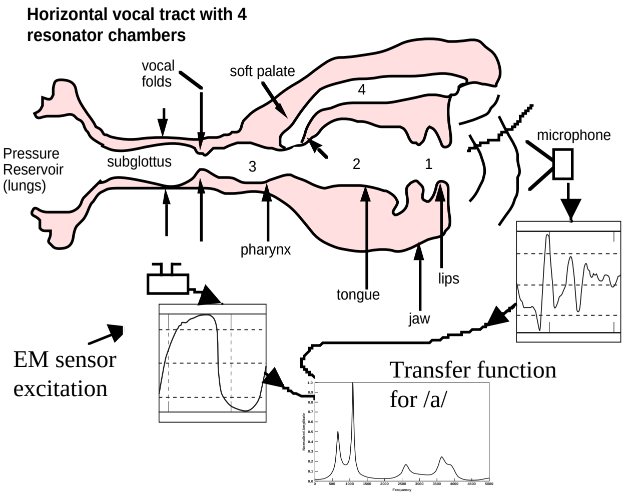

Horizontal vocal tract with 4 resonator chambers
vocal folds
soft palate
4
microphone
Pressure Reservoir (lungs)
subglottus
3
2
1
pharynx
tongue
lips
jaw
EM sensor excitation
Transfer function for /a/

Figure 1. A linearized vocal tract showing locations of EM sensors, corresponding excitation and acoustic functions, and resulting transfer function.

3. APPLICATIONS OF GEMS TO SPEECH DENOISING

The GEMS sensor is able to detect the transition boundaries between voiced and unvoiced or no speech. Because of the distinct differences in how voiced and unvoiced speech is produced, the methods of denoising are also different (4). GEMS measurements provide three advantages for speech processing. First, the GEMS signal can be used to define the onset of speech, end of speech, no-speech periods, and unvoiced periods. Second, filters built on the glottal signals can be used to suppress background noise that falls outside the pass-band of the excitation function. Third, the GEMS signal enables background noise spectral content to be determined during periods of no speech enabling an optimum "Correlation" filter to be built to maximally eliminate the background noise.

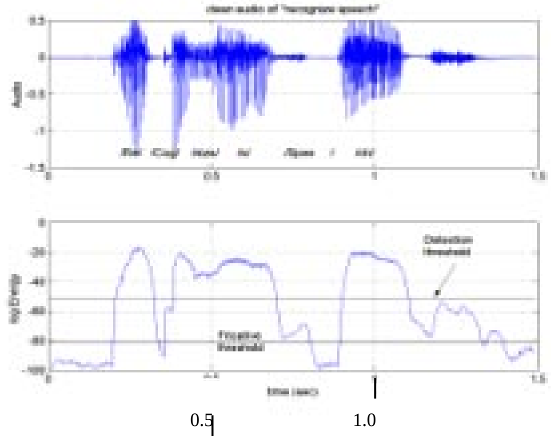


Figure 2. Clean acoustic and log acoustic energy versus time (sec.)

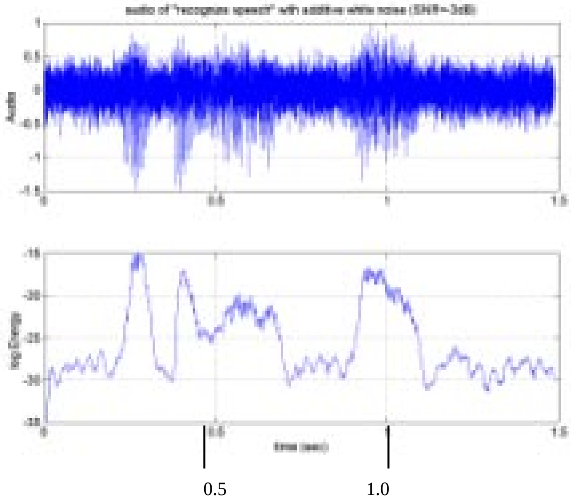


Figure 3. Noisy acoustic and log acoustic energy versus time (sec.)

For clean audio signals such as the words "recognize speech," shown in Fig. 2, the log of acoustic signal energy can be computed and thresholds can easily be chosen to separate voiced and unvoiced periods from the background noise. However, in a high noise environment such as that shown in Fig. 3, using the log of signal energy approach to differentiate speech boundaries. The boundaries become unreliable. In contrast, the GEMS signal, undisturbed by the background acoustics, remains a reliable means to measure voiced boundaries as shown in Fig. 4. By computing the time duration statistics (3) of unvoiced speech before (0.3 sec for American English) and after a voiced utterance (0.5 sec), one can statistically identify periods during which unvoiced speech is likely to occur. Further, time periods preceding or following the unvoiced time periods can be assumed to contain no speech, assuming there are no GEMS signals within appropriate time intervals. During no speech periods, accurate background noise spectral information can be measured and used to suppress background noise during a following or preceding speech period.

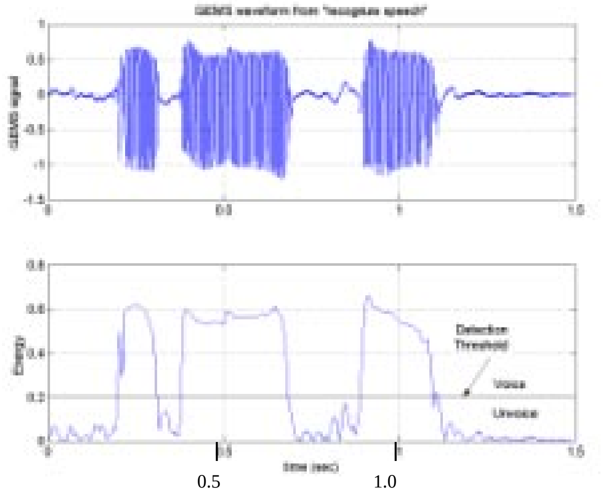
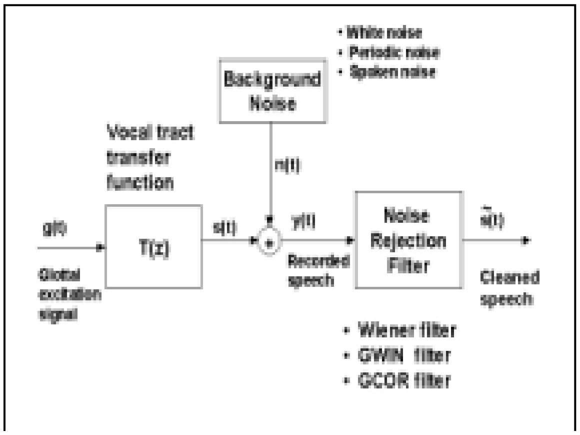


Figure 4. GEMS signal and energy versus time (sec.) corresponding to acoustic data in Fig. 2.



Three filtering methods, illustrated in Fig. 5, are employed to illustrate the noise suppression properties made possible by GEMS-like sensors: the Wiener filter, the glottal windowing (GWIN) filter, and the glottal correlation (GCOR) filter. The sequence of filters provides successively higher performance, relying on increasing knowledge of signal and noise statistics.

3A) WIENER FILTERING

The Wiener filter provides an optimum noise rejection capability when both the signal and noise spectrum is stationary and known. The Wiener filter is given by:

$$W(f) = \frac{P_{sy}(f)}{P_{yy}(f)} \cong \frac{P_{ss}(f)}{P_{ss}(f) + P_{nn}(f)} \qquad (1)$$

where Psy(f) and Pyy(f) are respectively the cross-power and auto-power spectral densities of the noise corrupted measurement and the desired signal, and Pnn(f) is the noise power spectrum. Note that Psy(f) reduces to Pss(f) when the signal and noise are uncorrelated. In general obtaining Pss(f) from a clean signal source is difficult, however, Pss(f) can be estimated from Pyy(f) by subtracting the noise portion Pnn(f) if available.

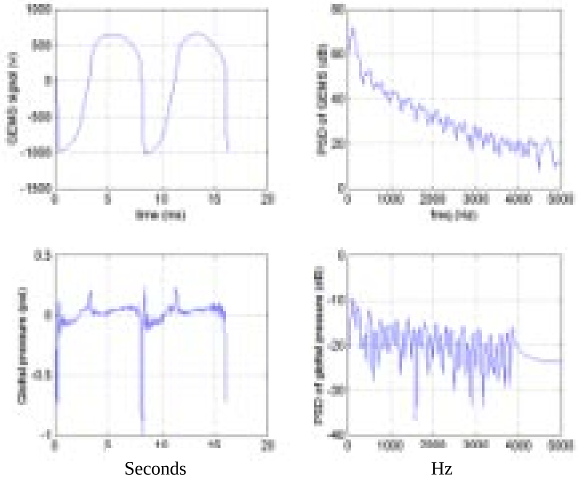


Figure 6. Glottal windowing (i.e., GWIN) windowing filter

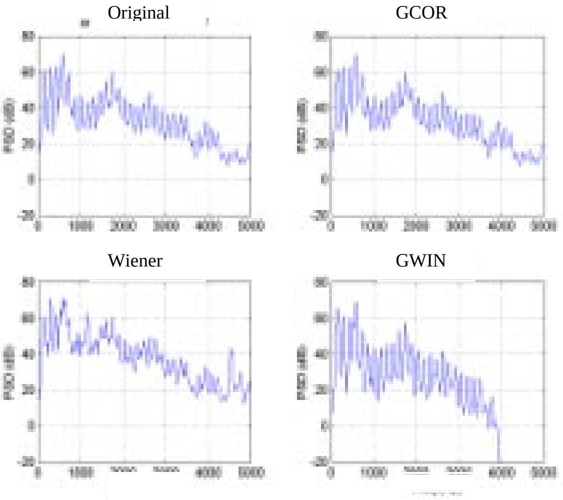


Figure 7. Original and 3 filters’ noise suppression performances. PSD vs Hz

3B) GWIN FILTERING

The glottal windowing (GWIN) filter is obtained by making use of the glottal excitation function as measured by the GEMS sensor. Since the audio signal s(t) is generated by the excitation function, then only harmonics within the window pass band of the excitation function are desirable. Background noise outside of the GWIN window will be suppressed. Of course, noise within the GWIN pass band will be retained. However GWIN is easy to construct and implement from the glottal signal. The GWIN filter is constructed by converting the glottal motion measurement to glottal air pressure (2). The air pressure signal is spectrally flatter, than the glottal signal as shown in Fig. 6.

3C) GCOR FILTERING

The glottal correlation filter can be constructed by making maximum use of information available. Since the measured audio signal can be written in the frequency domain as:

$$y(f) = T(f)\,G(f) + N(f) \tag{2}$$

where T(f) is the vocal tract transfer function. Multiplying both sides of Eq. (2) by G*(f), the complex conjugate, and taking the expectation, one obtains an estimate of the vocal tract transfer function as:

$$\tilde{T}(f) = \frac{P_{gy}(f) - P_{gn}(f)}{P_{gg}(f)} \tag{3}$$

and the cleaned audio signal can be generated from the equation:

$$\tilde{S}(f) = \tilde{T}(f)G(f) \tag{4}$$

Thus the concept of GCOR is to extract signal components that are correlated with the excitation function. Note that if the noise is uncorrelated with the glottal signal, i.e. Pgn(f) =0, the cleaned signal contains no noise at all. On the other hand, if Pgn(f) is not zero, then that component will be subtracted from Pgy(f) and again the cleaned signal contains no noise. Therefore the degree of knowledge of Pgn(f) directly determines the quality of the cleaned audio signal. However, when Pgn(f) is small compared to Pgy(f), it can be ignored without significantly affecting the quality of the reproduced audio signal.

The effectiveness of the three filters was compared using the simple open tube utterance /eh/ corrupted by additive noise (SNR~ -3dB). Fig. 7 shows the results as expected. The GCOR spectrum nearly matches the original signal, the Wiener filter is the worst since at low SNR, the Wiener filter is functioning principally to “whiten” rather than reject noise. The GWIN quality lies in the middle, rejecting most noise outside the pass band.

4. SPEECH DENOISING PROCESSING

Using the sentence “recognize speech” as a test example, background noise suppression is illustrated in Fig. 8. First the GEMS signal energy is used with a threshold to resolve the transition boundary between voiced and unvoiced period. For the voiced speech, a cleaned audio signal is constructed using either the GWIN or GCOR filter, over a two-glottal-cycle time period. The beginning and end times of the 2-glottal cycle period are used to process both the speech and GEMS signals to eliminate transient numerical effects. For the unvoiced speech, the Wiener filter is applied. The signal power spectrum, in the periods identified as likely for unvoiced speech, is estimated by removing the noise spectrum obtained from the no-speech time period. Fig. 8 shows example of the result of the reconstructed voiced and unvoiced speech segments which were corrupted by additive white noise. The cleaned speech essentially reproduces the original audio waveform, and sounds almost noise-free to a listener.

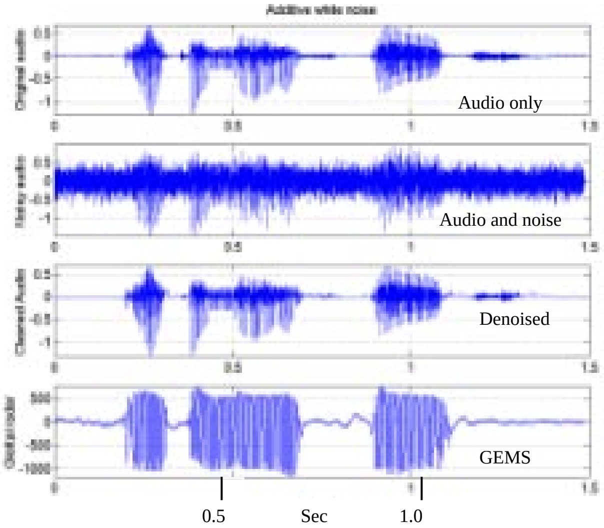


Figure 8. Example of background denoising from spoken word “recognize speech.” First graph is original acoustic signal, second graph is with –3dB white noise, third is denoised acoustic result, and fourth is GEMS signal.

5. CONCLUSION

Low power EM radar-like sensors can measure the internal properties of the human glottal regions safely and non-invasively. These data, together with the user’s speech signal, and reliable sampling of the acoustic noise signals enable several denoising algorithms to be employed that enable very understandable speech to be reconstructed, under conditions of S/N as low as 3dB.

ACKNOWLEDGMENTS

We thank the U.S. Department of Energy and the National Science Foundation for their support. Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

REFERENCES

[1] Holzrichter, J. F.; Burnett, G. C.; Ng, L. C.; and Lea, W. A., *Speech articulator measurements using low power EM-wave sensor,* J. Acoust Soc. Am. 103 (1) 622,1998. Also see the Web site http://speech.llnl.gov/ for related information.

[2] Burnett, G. C., *The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract,* Thesis UC Davis, Jan. 15th, 1999, document #9925723 available from University Microfilms, Inc. Ann Arbor, Michigan; also see Web site mentioned in [1].

(3) Herrnstein, A.; Holzrichter, J. F.; Burnett, G. C.; Gable, T. J.; and Ng, L., *Statistics of unvoiced time period duration relative to EM sensor detected voiced onset and end times* unpublished. Statistics are based upon a corpus of 15 male speakers pronouncing excerpts from a TIMIT phoneme, numeral, and sentence data set Corpus is contained on 8 CDs,
available as UCRL MI-132776.

(4) Rabiner, L. and Juang, B. W., *Fundamental of Speech Recognition,* Prentice-Hall, 1993.